Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado
Facultad de Ciencias Matemáticas
Unidad de Posgrado

Regresión no paramétrica mediante procesos de simulación

TESIS

Para optar el Grado Académico de Magíster en Estadística

AUTOR

Grabiela Yolanda MONTES QUINTANA

ASESOR

Antonio BRAVO QUIROZ

Lima, Perú

2019

Referencia bibliográfica

Montes, G. (2019). *Regresión no paramétrica mediante procesos de simulación.* [Tesis de maestría, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Hoja de metadatos complementarios

Código ORCID del autor	https://orcid.org/0000-0002-9234-7049
DNI o pasaporte del autor	08389428
Código ORCID del asesor	https://orcid.org/0000-0001-9051-2808
DNI o pasaporte del asesor	10130035
Grupo de investigación	G.I. MOCA
Agencia financiadora	Autofinanciado
Ubicación geográfica donde se desarrolló la investigación	Bellavista Callao *Coordenadas geográficas de Bellavista*, Latitud: -12.0625, Longitud: -77.1292 12° 3′ 45″ Sur, 77° 7′ 45″ Oeste.
Año o rango de años en que se realizó la investigación	Enero 2018 – diciembre 2019
Disciplinas OCDE	Estadísticas, probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03

ACTA DE SUSTENTACION DE TESIS DE GRADO ACADEMICO DE MAGISTER

Siendo las 18:00 p.m. horas del día jueves 12 de diciembre del dos mil diecinueve, en el Aula 201 de la Facultad de Ciencias Matemáticas, el Jurado Evaluador de Tesis, Presidido por el Dr. Helfer Joel Molina Quiñones e integrado por los siguientes miembros, Mg. José Antonio Cárdenas Garro (Jurado Evaluador), Mg. Carlos Alberto Jaimes Velásquez (Jurado Informante) y el Mg. Antonio Bravo Quiroz, como Miembro Asesor, se reunieron para la sustentación de la tesis titulada: «REGRESIÓN NO PARAMÉTRICA MEDIANTE PROCESOS DE SIMULACIÓN» presentada por la Bachiller Grabiela Yolanda Montes Quintana para optar el Grado Académico de Magister en Estadística.

Luego de la exposición de la graduanda, los Miembros del Jurado hicieron las preguntas correspondientes, así como las observaciones e inquietudes acerca del trabajo de tesis, a las cuales la Bachiller Grabiela Yolanda Montes Quintana respondió con acierto y solvencia, demostrando pleno conocimiento del tema.

A continuación, se realizó la calificación correspondiente, según tabla adjunta, resultando la Bachiller Grabiela Yolanda Montes Quintana aprobado con el calificativo de Muy bueno (18).....

Habiendo sido aprobada la sustentación de la Tesis, el Jurado Evaluador recomienda para que el Consejo de Facultad apruebe el otorgamiento del Grado Académico de Magister en **Estadística** a la Bachiller Grabiela Yolanda Montes Quintana.

Siendo las 18:55 horas, se levantó la sesión, firmando para constancia la presente Acta.

Mg. José Antonio Cárdenas Garro
MIEMBRO

Mg. Helfer Joel Molina Quiñones
PRESIDENTE

Mg. Carlos Alberto Jaimes Velásquez
MIEMBRO

Mg. Antonio Bravo Quiroz
MIEMBRO ASESOR

El presente trabajo lo dedico a mis padres, a quienes todavía tengo a mi lado, a mi querido esposo Mg. Wilfredo Domínguez por su aliento y apoyo constante, finalmente a mis hijos.

Agradezco a todos los docentes de la unidad de posgrado de la facultad de ciencias matemáticas, especialmente a mi asesor Mg. Antonio Bravo.

Índice

RESUMEN

REGRESIÓN NO PARAMÉTRICA MEDIANTE PROCESOS DE SIMULACIÓN

Presentado por: Lic. MONTES QUINTANA, Grabiela Yolanda

Profesor Asesor: Magister BRAVO QUIROZ, Antonio

JUNIO 2019

En el presente trabajo de investigación se hace una revisión preliminar de los modelos de regresión, presentando los modelos de regresión lineal simple y múltiple, conocidos como paramétricos. Luego se estudian los modelos de regresión no paramétricos, describiendo los más importantes. Finalmente se realiza un estudio de simulación para comparar dos modelos de regresión no paramétricos, Loess y suavización por Splines. Se utilizan diferentes funciones de regresión, como son funciones sin oscilaciones, con pocas oscilaciones y con muchas oscilaciones, también se utilizan diferentes distribuciones para el término de error del modelo de regresión, estos son simétricas, asimétricas hacia la izquierda y asimétricas hacia la derecha. Del estudio se obtiene como resultado un buen comportamiento del modelo de regresión no paramétrico por Splines, para los casos de modelos sin oscilaciones o con muchas oscilaciones. Concluyendo que en el caso de modelos con pocas oscilaciones ambos métodos son igualmente eficientes.

Palabras clave: Regresión no paramétrica, Kernel, suavización, Loess, Splines.

ABSTRACT

In the present work of investigation a preliminary revision of the models of regression is made, presenting the models of linear regression simple and multiple, known like parametric. Then the nonparametric regression models are studied, describing the most important ones. Finally, a simulation study is carried out to compare two non-parametric regression models, Loess and Splines smoothing. Different regression functions are used, such as functions without oscillations, with few oscillations and with many oscillations, different distributions are also used for the error term of the regression model, these are symmetric, asymmetric to the left and asymmetric to the right. From the study, a good performance of the nonparametric regression model is obtained by Splines, for the cases of models without oscillations or with many oscillations. Concluding that in the case of models with few oscillations both methods are equally efficient.

Keywords: Nonparametric regression, Kernel, smoothing, Loess, Splines.

CAPÍTULO 1: INTRODUCCIÓN

1.1 Situación Problemática

El término regresión fue utilizado por primera vez por Francis Galton en el siglo XIX, en un estudio donde este postuló la hipótesis de que los hijos de padres altos tienden a ser altos, pero no tanto como sus padres y, los hijos de padres bajos tienden a ser bajos pero no tanto como sus padres, ya que las estaturas de los hijos tendían a regresar a la estatura promedio de la población a la que pertenecían. Dentro del campo de la modelización estadística generalmente se estudian situaciones en las cuales se cuenta con una o más variables independientes cuantitativas, y se quiere saber si a partir de un modelo llamado modelo de regresión, se puede predecir el valor de otra variable llamada dependiente.

Modelar estadísticamente consiste en separar cada valor que toma la variable Y, dependiente, de cada individuo, en dos elementos, uno función de las variables independientes o explicativas, y otro que es propio del individuo como se muestra a continuación:

$$y_i = f\left(x_{i1},\ldots,x_{ip}\right)+\varepsilon_i$$

donde f es la función que relaciona los valores de la variable dependiente con las independientes, mientras que ε_i es una variable aleatoria, que corresponde solamente al individuo i que no está explicada por ninguna otra variable.

La función f representa la parte determinista del modelo, que permite explicar cómo se comporta la variable dependiente y hace posible poder predecir valores, mientras que ε_i representa la parte impredecible, aleatoria y se denomina término de error. Este término de error tiene una distribución de probabilidad que permite obtener los distintos valores para cada individuo, se puede suponer que su valor esperado es cero.

El problema aparece cuando la función f no puede ser especificada en función de parámetros, y por lo tanto no podemos utilizar esta función para realizar las predicciones indicadas. Se hace pues necesario utilizar otras técnicas que no necesiten del conocimiento específico de esta función. Los modelos llamados paramétricos presentan pues el problema de tener una estructura muy rígida, poco flexible, y de difícil adaptación en situaciones complejas, como una alternativa aparecen los modelos de regresión llamados no paramétricos, que suponen solamente que la función f sea derivable, es decir que tenga una estructura suave de manera que se puedan realizar las predicciones de los valores de la variable dependiente, a partir de los valores de observaciones cercanas al punto que se quiere predecir. Como principal ventaja los métodos de regresión no paramétrica presentan su flexibilidad, aunque requiere desde el punto de vista teórico una mayor complejidad y un mayor coste computacional.

1.2 Formulación del Problema

¿Qué métodos de regresión no paramétrica son adecuados para predecir valores de la variable dependiente?

1.3 Justificación de la Investigación

Para la aplicación de los métodos de regresión clásica, donde se conoce la forma de la función de regresión pero se desconocen los valores de los parámetros de esta función, para realizar la predicción de los valores de la variable dependiente, se estiman los valores de los parámetros, aplicando los métodos conocidos, como mínimos cuadrados o máxima verosimilitud, y así utilizando estos estimadores se obtiene el estimador de la función de regresión, pero cuando la forma de la función de

regresión es desconocida, se debe encontrar una alternativa para solucionar el problema.

En el estudio de los métodos de regresión no paramétrica se utiliza la información contenida en los datos aplicando métodos de suavización para obtener una estimación de la verdadera función de regresión.

Los métodos de regresión no paramétrica, se han desarrollado de manera amplia en los últimos tiempos, por el avance de las computadoras, haciendo posible la implementación de éstos, ya que para su aplicación se necesita cálculo intensivo, entonces en el presente trabajo, se busca conocer y comparar algunos de los métodos de regresión no paramétricos y obtener los programas computacionales que ayuden a su aplicación en casos reales.

Los resultados de este trabajo de investigación podrán motivar a otros investigadores a continuar con el estudio de este tema.

1.4 Objetivos de la Investigación

1.4.1 Objetivo General

Comparar los métodos de regresión no paramétrica, mediante un estudio de simulación para diferentes situaciones.

1.4.2 Objetivos Específicos

- Comparar los métodos de regresión no paramétrica, estimadores de Kernel, LOESS y estimación Spline.
- Aplicar el software R para realizar el estudio de simulación.

CAPÍTULO 2: MARCO TEÓRICO

2.1 Antecedentes del Problema

- En 1988, Randall R. Eubank, presenta su trabajo Non Parametric Regression and Spline Smoothing, libro editado por Marcel Dekker Inc. New York, el cual trata de forma completa el tema de regresión no paramétrica en forma teórica.
- En el 2001, Luca Scrucca del Departamento de Estadística de la Universidad degli Studi di Perugia, Italy, presenta un trabajo titulado Nonparametric Kernel Smoothing Methods. The sm library in Xlisp.Stat, en el cual describe la utilización del software Xlisp.Stat para aplicar los métodos de suavización de Kernel, en el cual también podemos encontrar una descripción completa de los métodos estadísticos utilizados. En este trabajo se puede encontrar métodos para estimar la función de densidad y la función de regresión.
- En el 2011, John Fox & Sanford Weisberg presentaron su trabajo, Non Parametric Regression in R, An Appendix to An R Companion to Applied Regression, Second Edition. En este trabajo se describe la aplicación del software estadístico R a la regresión no paramétrica, pero también a modelos de regresión múltiple, modelos de regresión aditivos y modelos de regresión no paramétricos generalizados.
- En la Revista EIA, ISS 1794-1237 Número 18, p 19-31, diciembre 2012, Jhovana Reina y Javier Olaya, presentaron su trabajo “Ajuste de Curvas Mediante Métodos No Paramétricos para Estudiar el Comportamiento de Contaminación del Aire por Material Particulado PM10”, en el cual concluyen que uno de los principales factores que contribuyen a la contaminación del aire es el material particulado de diámetro aerodinámico inferior a 10 micrómetros,

conocido como PM10. Como la presencia de este contaminante varía de forma irregular y temporal en la atmósfera, era necesario caracterizar un modelo de suavización no paramétrico. Como resultado encontró que la variable PM10 tiene curvas estimadas unimodales durante las horas de la mañana en el norte de Cali Colombia, teniendo diferente comportamiento los distintos días de la semana y en los días con lluvia y sin lluvia.

- En el volumen 55 de la revista Journal of Statistical Software, octubre 2013, Issue 2, Kris De Brabanter, Johan A. K. Suykens y Bart De Moor, presentan su trabajo 'Nonparametric Regression via Stat LSSVM', en el cual se desarrollan herramienta nuevas MATLAB para Window y Linux para estimación de regresión no paramétrica, basada en la librería estadística AtatLSSVM,
- En la revista EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS, vol. 6, N° 2, 2013, pp 222-238, se publicó el trabajo 'Smoothing Parameter Selection for Nonparametric Regression Using Smoothing Spline', escrito por Dursun Aydin, Memmedaga Memmedli y Rabia Ece Omay. Este trabajo se enfoca en la selección del parámetro de suavización respecto a la suavización Spline, en la predicción de modelos de regresión no paramétrica.
- En 'Comunicaciones en Estadística', Junio 2014, Vol 7, N° 1, pp 49-66, Álvaro José Flórez y Javier Olaya, presentan su trabajo "Estudio de simulación para comparar varios estimadores de varianza en el marco de la regresión no paramétrica", en este trabajo se presentan diferentes estimadores de la varianza los cuales utilizan métodos de diferencias, en regresión no paramétrica. La ventaja de estos estimadores es que son independientes de los parámetros de suavización. Usaron métodos basados en diferencias ordinarias y en diferencias óptimas de Hall, aplican estos métodos a casos de distribuciones asimétricas de los errores. Los resultados apoyaron

el supuesto de que los estimadores basados en diferencias óptimas de Hall no son mejores en todos los casos estudiados.

- En 2013, Luis Meza presentó la tesis “Regresión No Paramétrica utilizando Spline para la suavización de la estructura de la mortalidad en el Perú”. En el cual se realizó un estudio preliminar del análisis de regresión en general y de la regresión no paramétrica en particular. Utiliza la regresión Spline para suavizar la curva que describe la estructura de mortalidad por sexo y edad. Utilizan la información de las defunciones de las estadísticas vitales del año 2007 y del Censo Nacional de Población del 2007 del departamento de Lima.

2.2 Bases Teóricas

2.2.1. Modelo de regresión paramétrica

En este caso, la regresión lineal simple, es el modelo de regresión más utilizado. Un modelo lineal general puede escribirse como:

$$y = f(x) + \varepsilon \tag{2.1}$$

donde *y* es una variable aleatoria, f es una función de una variable no-aleatoria definida en un dominio D. La función de regresión f es la parte determinística del modelo. La variable aleatoria ε es no-observable, se asume que esta variable tiene una función de densidad, generalmente con media cero.

Una característica de estos modelos de regresión es que se conoce la forma de la función de regresión f, y esta función tiene parámetros desconocidos. Se dice que el modelo es lineal porque f es una función lineal de los parámetros desconocidos.

La función de regresión lineal f más simple es

$$f(x) = \beta_0 + \beta_1 x \tag{2.2}$$

Para $x \in D_x$; β_0 y β_1 se definen en un espacio paramétrico Ω_β.

Si reemplazamos (2.2) en (2.1) tenemos:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.3}$$

donde $E(\varepsilon) = 0$ y $V(\varepsilon) = \sigma^2$. Generalmente σ^2 es desconocido. El modelo 2.3 es conocido como modelo lineal simple.

Con el análisis del modelo lineal simple se busca estimar β_0 y β_1. También se necesita estimar σ^2, la varianza del error, para poder llevar a cabo las inferencias a partir del modelo 2.3.

El modelo dado en 2.3 se conoce habitualmente como *modelo poblacional*. Para estimar los parámetros, se necesita un *modelo muestral*, el cual puede escribirse como el conjunto de ecuaciones

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1,...,n$$
$$\text{con} \quad E(\varepsilon_i) = 0 \tag{2.4}$$

donde (x_i, y_i), i = 1,…n son observaciones de la variable X y de la variable Y. Los ε_i son variables aleatorias no-observables tales que $cov(\varepsilon_i, \varepsilon_j) = \sigma_{ij}$.

El modelo lineal general 2.1 se puede extender al caso de una covariable q-dimensional y se puede escribir como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2,..., n.$$

En la expresión anterior q = p + 1.

En este caso es más práctico escribir el modelo en la notación matricial

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.5}$$

donde **y** es un vector aleatorio observable de orden n x 1; **X** es una matriz n x q de números dados (los elementos de X no son variables aleatorias); **β** es un vector q x 1 de parámetros no-observables definidos en un espacio paramétrico Ω, y **ε** es un vector aleatorio n x 1 tal que E(**ε**) = **0** y cov(**ε**) = ∑, siendo ∑ una matriz definida positiva, q = p + 1 en general y en tal caso la primera columna de **X** tiene todos sus elementos iguales 1 y el vector **β** tiene los elementos $\beta_0, \beta_1, \cdots \beta_p$. Las restantes p columnas de **X** contienen las observaciones del vector **x** de variables continuas independientes.

Para el caso particular de **x** = (X_1, X_2), tenemos lo siguiente:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2.6}$$

Para estimar la función de regresión f en el modelo lineal, se puede utilizar el método conocido como de los mínimos cuadrados ordinarios.

Los estimadores de los parámetros β_j de la expresión 2.4, de mínimos cuadrados están dados por:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.7}$$

donde $\bar{x}$ y $\bar{y}$ son respectivamente la media aritmética de X y de Y, definidas como:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Luego, un estimador $\hat{f}(x)$ de la función de regresión $f(x)$ en el modelo lineal simple 1.4 estaría dado por

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Para el caso del predictor q-dimensional y si $(\mathbf{X}^T\mathbf{X})^{-1}$ existe, entonces el estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}$ del vector de parámetros $\boldsymbol{\beta}$ está dado por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.8}$$

Se demuestra que $\hat{\boldsymbol{\beta}}$ es un estimador insesgado de $\boldsymbol{\beta}$ ya que se ha supuesto que E($\boldsymbol{\varepsilon}$) = **0,** por lo tanto $\mathbf{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ de donde se obtiene que $\mathbf{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{E}(\mathbf{y}) = \boldsymbol{\beta}$, y luego se concluye que $\hat{\mathrm{f}}(\mathrm{x})$ es un estimador insesgado de f(x), es decir $\mathrm{E}\left[\hat{\mathrm{f}}(\mathrm{x})\right] = f(x)$.

Además como $\mathrm{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$, se verifica que:

$$\mathrm{V}(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\sigma^2 = \mathbf{H}\sigma^2$$

Lo que nos indica que el valor de V(f) depende de los valores observados de **x**.

En el caso de que $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$ sea singular, entonces no tendríamos una solución única para $\hat{\boldsymbol{\beta}}$. Esto significaría que al menos una columna de X es linealmente dependiente de una o más de las otras columnas. Esto conduce a estimaciones poco confiables de los parámetros, ya que estos tendrán varianzas y covarianzas grandes (Draper & Smith 1998).

También se podría ajustar un modelo usando polinomios de grado p de la forma:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i, \qquad i = 1,2,...,n,$$

Que equivale a escribir

$$y_i = \sum_{i=0}^{p} \beta_j x_i^j + \varepsilon_i \tag{2.9}$$

Este problema se resuelve usando mínimos cuadrados ordinarios, con la matriz **X** de la forma:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix}$$

Un caso más general del modelo (2.1) se presenta cuando se considera una covariable aleatoria X. Aquí se asume que disponemos de dos variables aleatorias W y X cuya función de densidad conjunta es $\Psi_{W,X}(w,x)$. Sea Y la variable aleatoria [W|X = x], E(ε) = 0 y V(ε) = $\sigma^2 < \infty$. Diremos que f(x) es el valor esperado de la variable aleatoria [W|X = x] y por lo tanto el modelo (2.1) estaría dado por:

$$Y = E(W \mid X = x) + \varepsilon \qquad (2.10)$$

Cuando se asume el modelo (2.10), ya no se necesita el modelo muestral (2.4), ya que en este caso estamos muestreando una densidad conjunta. Finalmente, podemos suponer que tenemos un vector aleatorio p-dimensional **x** como covariable. En este caso asumiremos que disponemos de p + 1 variables aleatorias $X_0, X_1,\ldots,X_p$, definiremos Y = $[X_0 \mid X_1,\ldots,X_p]$ y asumiremos que la función de regresión f(**x**) está dada por el valor esperado de la variable aleatoria $[X_0 \mid X_1 = x_1,\ldots,X_p = x_p]$. Luego recolectamos n observaciones $\{x_{0i}, x_{1i}, x_{2i},\ldots,x_{pi}\}$, i = 1,…,n, de la densidad conjunta $\Psi_{X_0,X_1,\ldots,X_p}(x_0, x_1,\ldots,x_p)$ y asumiremos que el modelo

$$Y = f(X_1,\ldots,X_p) + \varepsilon$$

se satisface para las variables aleatorias Y, X_1, $X_2,\ldots,X_p$ y ε. Además se supone que E(ε) = 0 y V(ε) = $\sigma^2 < \infty$.

Finalmente, la distribución de Y = $[X_0 \mid X_1,\ldots,X_p]$ satisface que

$$E[X_0 \mid X_1 = x_1,\ldots,X_p = x_p] = f(x_1,x_2,\ldots,x_p)$$

Por lo que la función de regresión será el valor esperado de la variable aleatoria condicional $[X_0 \mid X_1 = x_1,\ldots,X_p = x_p]$.

Para el caso de los estimadores dados en (2.7), se obtiene $\hat{f}(x)$, el cual es un estimador insesgado de f(x), asumiendo que la verdadera función de regresión f es lineal. Además, la varianza de $\hat{f}(x)$ está dada por

$$V\left(\hat{f}(x)\right) = \sigma^2\left[n^{-1} + (x-\bar{x})^2 \Big/ \sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

donde σ^2 es la varianza de los errores del modelo 2.3.

Por otra parte, $\hat{f}$ es un estimador consistente de f, cuyo error cuadrático medio converge a cero a una tasa de n^{-1}. En consecuencia, podemos concluir que $\hat{f}$ es un buen estimador de f.

2.2.2. Regresión No Paramétrica

En la regresión no paramétrica simple donde se utiliza un solo predictor, tenemos n observaciones bivariadas, denotadas $(x_1, y_1), (x_2, y_2),\ldots,(x_n, y_n)$, de la variable respuesta Y para n valores conocidos de una variable independiente X. Las n observaciones bivariadas disponibles, siguen el modelo

$$y_i = f(x_i) + \varepsilon_i \qquad i = 1,...,n \tag{2.11}$$

donde f es una función de regresión desconocida y se satisface que , $\varepsilon = (\varepsilon_1,...,\varepsilon_n)^T$ es un vector de errores aleatorios no correlacionados que tienen media cero y varianza común σ^2.

La regresión no paramétrica simple también es conocida como *scatterplot smoothing* (Suavizadores de diagramas de dispersión).

Sin pérdida de generalidad asumiremos que los valores de X han sido elegidos de la siguiente manera:

$$x_i = (2i-1)/2n, \qquad i = 1,...,n \tag{2.12}$$

Se quiere estimar f en (2.11), para lo cual se construirán estimadores lineales, que para un λ dado, es una combinación lineal de las observaciones y_i, donde $K(.,x_i;\lambda)$, i = 1,…,n es una colección de funciones ponderadoras que dependen de los x_i y de un parámetro de suavización denotado λ. Los estimadores lineales serán de la forma:

$$f_\lambda(x) = \sum_{i=1}^{n} K(x, x_i; \lambda)\, y_i \tag{2.13}$$

Entre los modelos univariados más utilizados tenemos:

- El regresorgrama desarrollado por Tukey (1961).
- Promedios móviles (Running means), Medianas móviles (Running medians), Líneas móviles (Running lines).

- Método de suavización de Kernel, desarrollado por Nadaraya-Watson, (1964).
- Regresión Local Ponderada, suavizado de diagrama de dispersión estimado localmente, LOESS, estudiado por Cleveland (1979).
- Método de Regresión No Paramétrico Spline, por Stone y Koo (1985).
- Suavización por Splines por Wahba (1975).

En el caso de múltiples variables predictoras tenemos los siguientes modelos de regresión no paramétrica

- Regresión de búsqueda de proyección (Regresión por Projection Pursuit), PPR desarrollado por Friedman, Stuelze, (1981).
- Regresión por árboles CART, por Breiman, Friedman, Olsen y Stone, (1984).
- Modelos Aditivos Generalizados por Hastie y Tibshirani (1990).

Se presentará a continuación una breve descripción de los métodos mencionados.

2.2.2.1 Regresorgrama

Acuña E. (2007) afirma lo siguiente:

> Aquí se divide los valores de la variable predictora en varios subintervalos (usualmente 5). La amplitud de los subintervalos se

> elige de tal manera que haya aproximadamente igual número de datos en cada uno de ellos. Luego se promedia los valores de la variable respuesta en cada subintervalo. Esto determina varios segmentos de línea que al unirse forma el regresorgrama. Lo malo de este estimador es que no es suave porque hay saltos en cada punto de corte (p. 214).

En la figura 1 presentamos un ejemplo del regresorgrama correspondiente a 100 puntos, donde la variable X es la secuencia de puntos del 1 al 100 y la Y se ha generado con la función:

$$Y(i) = 5*x(i) + rnorm(1, 50, 100)$$

donde rnorm(1, 50, 100) es un valor aleatorio de la distribución normal con media 50 y desviación estándar 100, en el programa R project.

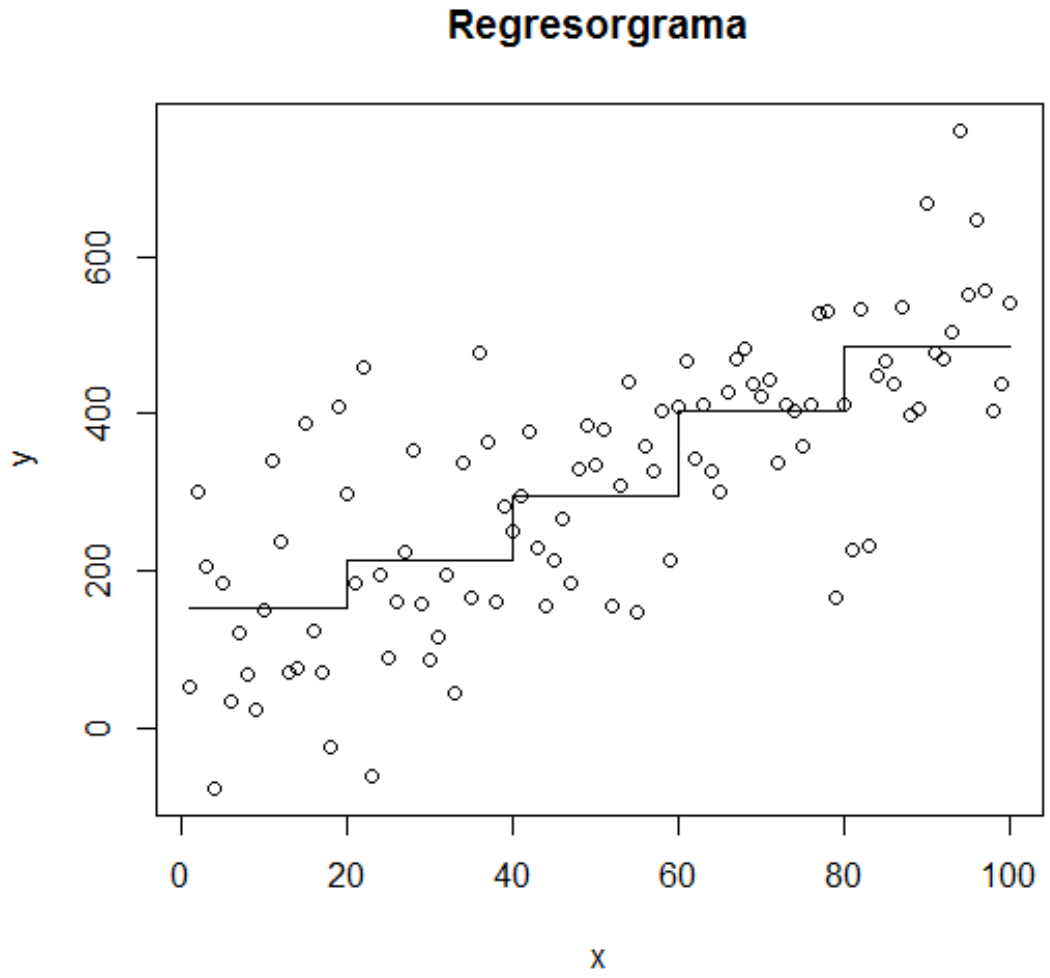

Figura 1. Gráfico de la suavización usando regresorgrama con 100 puntos generados con una función de regresión dada.

2.2.2.2 Promedios móviles (Running means), Medianas móviles, Líneas móviles:

Acuña E. (2007) afirma lo siguiente:

Obtenemos un entorno correspondiente a cada xi, el cual se denominará N(xi), este entorno debe contener al punto xi y el mismo número de puntos

x a la derecha como a la izquierda de xi. Esta condición no se podrá lograr en los extremos, pero se buscará acercarse lo más posible, a esta condición se le conoce como vecindad simétrica.

Luego se obtiene el estimador por suavización por “promedios móviles” en el punto xi, calculando el promedio de los y correspondientes a las x que caen en N(xi).

En el caso de medianas móviles, el suavizador se calcula utilizando la mediana, y en el suavizador por “líneas móviles”, se calcula para x = xi, el valor estimado de la regresión mínimo cuadrática que se obtiene usando los puntos (xi, yi) con los x que caen en N(xi).

A continuación se presenta el estimador por medias móviles para el ejemplo presentado anteriormente:

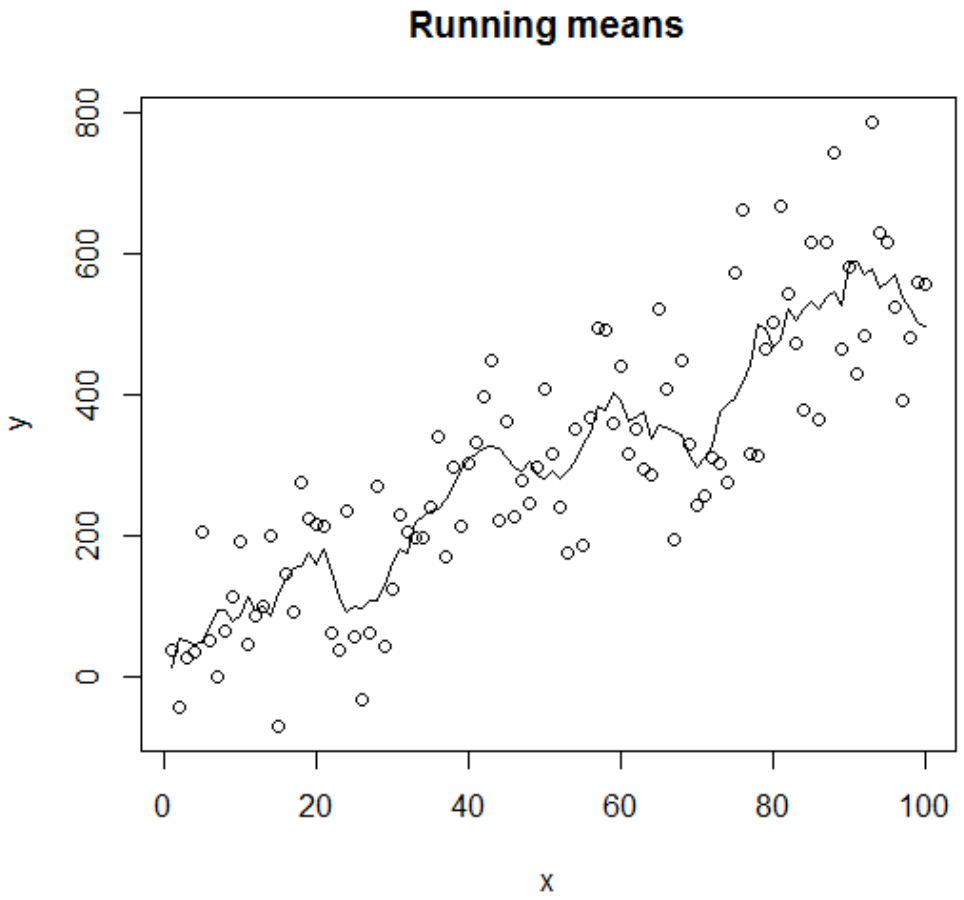

Figura 2. Gráfico de la suavización por Medias Móviles de los datos del ejemplo.

2.2.2.3 Suavización por Kernels:

Siguiendo a Acuña E. (2007), encontramos lo siguiente:

Considerando que tanto x como y son aleatorias se puede escribir $g(x) = E(y/x) = \int y\ f(y/x)\, dy$ donde f(y/x) representa la función de densidad condicional de y dado x. Usando la definición de densidad condicional lo anterior se puede re-escribir como

$$g(x) = \frac{\int y\, f(x, y)\, dy}{f(x)}$$

Las funciones f(x) y f(x, y), son estimadas usando suavización por Kernel, a partir de los datos (x_i, y_i) de la muestra.

Más específicamente

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h} \right)$$

y

$$\hat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h} \right) K\left(\frac{y - y_i}{h} \right)$$

donde K(t) es un Kernel, el cual es una función no negativa, simétrica con respecto a 0 y con valor máximo en dicho punto. Además $\int_{-\infty}^{\infty} K(t)\, dt = 1$.

Las funciones de Kernel asignan pesos más bajos a observaciones más alejadas del punto de estimación y al mismo tiempo, asigna pesos altos a las observaciones cercanas al punto de estimación.

Estos estimadores se obtienen utilizando las siguientes funciones de ponderación:

$$K(x, x_i; \lambda) = \frac{1}{\lambda} K\left(\frac{x - x_i}{\lambda} \right) \qquad i = 1, \ldots, n \tag{2.14}$$

donde K es una función simétrica con soporte en [-1,1] tomando su máximo valor en cero. En este caso el parámetro λ, es un número real no negativo, llamado parámetro de suavización o ancho de ventana. Es de dos tipos, fijo

o variable. Es fijo cuando el valor de λ es el mismo para todos los valores de x, es variable, como ocurre en el método de los k-vecinos más cercanos, cuando para cada valor de x, se utilizan la misma cantidad de vecinos más cercanos, pudiendo ser que estos estén contenidos en intervalos de diferentes tamaños. Si se utilizan longitudes de ventana fijas, el sesgo del estimador permanece constante, pero la varianza se comporta inversamente proporcional a la densidad local. En cambio para ventanas de k-vecinos más cercanos la varianza se mantiene constante pero el sesgo varía con la densidad local.

K es una función simétrica con soporte en [-1,1] tomando su máximo valor en cero. En este caso el parámetro λ, es un número real no negativo, llamado parámetro de suavización o ancho de ventana

$$\int_{-1}^{1} K(u)\,du = 1$$
$$\int_{-1}^{1} u\ K(u)\,du = 0 \qquad (2.15)$$

La funciones K se denominan *funciones Kernel*, y a los estimadores basados en estas funciones se denominan *estimadores Kernel*.

A continuación se presentan los Kernels más utilizados en su forma unidimensional. Todos tienen su versión multidimensional que se pueden encontrar muy fácilmente, por ejemplo en el caso de dos dimensiones K(u, v) = K(u) K(v).

- Kernel rectangular o uniforme: La utilización de éste lleva al método de histogramas móviles.

$$K(u) = \frac{1}{2} I_{[-1,1]}(u)$$

- Kernel Triangular: Se denomina así por su forma gráfica.

$$K(u) = (1 - |u|\ ) I_{[-1,1]}(u)$$

- Kernel Gaussiano: Se llama de esta manera al ser la densidad de una variable aleatoria normal estándar.

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

- Kernel de Epanechnikov o cuadrático: Aunque la elección del Kernel no es tan importante como el valor de λ, éste es con el que se obtienen mejores resultados.

$$K(u) = \frac{3}{4} (1 - u^2) I_{[-1,1]}(u)$$

- Kernel Biponderado:

$$K(u) = \frac{15}{16} (1 - u^2)^2\ I_{[-1,1]}(u)$$

En la figura 3 se presenta los gráficos de algunas funciones Kernel

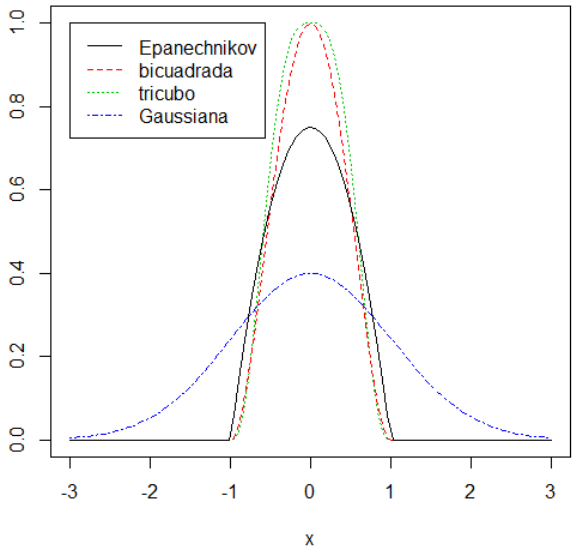

Figura 3. Se presenta los gráficos de algunas funciones Kernel

El parámetro λ permite controlar la cantidad de observaciones que participarán en la estimación. La elección de λ es muy importante en la estimación no paramétrica, ya que cambios de este valor influyen considerablemente en la estimación de la función de regresión.

Estimador de kernel de Priestley-Chao

Este estimador fue propuesto por Priestley & Chao (1972) y tiene la forma

$$f_\lambda(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right) y_i \qquad (2.16)$$

A este estimador se le conoce también como Kernel genérico.

El parámetro de suavización, λ, se elige utilizando el criterio de validación cruzada generalizada GCV, el cual será presentado posteriormente.

Estimador de Nadaraya-Watson

Como una alternativa para mejorar el comportamiento del estimador Kernel genérico, Benedetti (1975), propone un estimador para un diseño de puntos de la forma

$$x_i = (2i-1)/2n, \qquad i = 1,...,n$$

Este estimador primero fue propuesto por Nadaraya & Seckler (1964) y Watson (1964), para el caso que X es una variable aleatoria. Este estimador se conoce como estimador de Nadaraya-Watson y se define de la siguiente manera:

$$f_\lambda(x) = \frac{\sum_{i=1}^{n} K\left(\lambda^{-1}(x - x_i)\right) y_i}{\sum_{j=1}^{n} K\left(\lambda^{-1}(x - x_j)\right)} \tag{2.17}$$

Este estimador realmente es un promedio ponderado de las observaciones cercanas al punto de observación.

Estimador de Gasser-Müller

Este estimador kernel fue propuesto por Gasser & Müller (1979), y está dado por:

$$f_\lambda(x) = \sum_{i=1}^{n} \left(\int_{s_{i-1}}^{s_i} K\left(\lambda^{-1}(x-s)\right) ds \right) y_i \tag{2.18}$$

donde

$$s_0 = 0, \quad s_{i-1} \le x_i \le s_i, \quad i = 1,..., n-1, \quad s_n = 1$$

Los tres estimadores kernel presentados, Priestley-Chao, Nadaraya-Watson y Gasser-Müller, tienen aproximadamente las mismas propiedades asintóticas.

2.2.2.4. Regresión Local Ponderada

Uno de los métodos más utilizados de la regresión polinomial local es el método Loess desarrollado por Cleveland (1979).

La estimación de Kernel, calcula promedios localmente ponderados, en cambio Loess estima funciones de regresión ponderando las observaciones en las proximidades de distintos focos.

Al estimador de kernel obtenido de minimizar la suma siguiente:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \cdots - \beta_p (x - x_i)^p \right)^2 K\left(\frac{x - x_i}{\lambda} \right) \tag{2.19}$$

se le conoce como estimador polinomial de regresión local de orden p.

La expresión más simple de (2.19) está dada por:

$$\sum_{i=1}^{n}(y_i - \beta_0)^2 K\left(\frac{x - x_i}{\lambda}\right) \tag{2.20}$$

Al resolver del sistema (2.20) obtenemos el estimador conocido como estimador LOESS, en este método se consideran los c vecinos más cercanos y el estimador de la función de regresión en el punto x, el procedimiento, según Olaya (2012), se presenta a continuación:

1. Considerando al valor x, se obtienen los c vecinos más cercanos a este valor, a este conjunto se denota como N(x).
2. Se calcula la distancia de x a cada punto dentro de N(x), y a la mayor distancia se le denota por Δ(x).

$$\Delta(x) = \max_{\{xi \in N(x)\}} |x - x_i|$$

3. Usando la función de pesos tri-cubo, se asignan los pesos w_i a cada punto en N(x)

$$W\left(\frac{|x - x_i|}{\Delta(x)}\right)$$

donde

$$W(u) = \begin{cases} (1 - u^3)^3, & si \quad 0 \le u \le 1 \\ 0 \quad , & c.c. \end{cases}$$

4. Considerando al conjunto N(x) y utilizando los pesos dados, se ajusta una recta por mínimos cuadrados ponderados de Y sobre X.

La estimación LOESS de la función de regresión $f_\lambda(x)$ en el punto x, está dada por el término independiente de la recta de regresión local resultante.

Este método puede ser generalizado a varios predictores, pero a partir de 3 o 4 predictores aparece el problema de la dimensionalidad, el cual consiste en encontrar muy pocas observaciones cercanas al punto en consideración, lo cual reduce la capacidad del modelo.

2.2.2.5. Regresión por Splines

Se sabe que la regresión polinomial tiene recursos limitados con respecto a la naturaleza global del ajuste mientras que los suavizadores tienen una naturaleza local explícita. La regresión por splines es más flexible que un modelo polinomial y tienen menos probabilidad de presentar multicolinealidad cuando se utilizan altas dimensiones. Se tiene también que los métodos splines son generalmente más eficientes que los métodos Kernel.

Los splines se implementaron en estadística por Wabba en 1990, pero sus orígenes se remontan a 1923 cuando Whittaker desarrolla la teoría. Un spline es simplemente una curva. En matemática, un spline es una función especial definida por polinomios.

Consideremos de nuevo el modelo

$$y_i = f(x_i) + \varepsilon_i \qquad i = 1,...,n$$

donde f es una función de regresión desconocida y se satisface que , $\varepsilon = (\varepsilon_1,...,\varepsilon_n)^T$ es un vector de errores aleatorios no correlacionados que tienen media cero y varianza común σ^2.

Una manera de estimar la función f(x) consiste en minimizar la siguiente suma de cuadrados:

$$\min_{\hat{f}:R\to R} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2$$

Este problema se soluciona aplicando métodos numéricos, como lo es la obtención de una función para interpolar los datos.

Generalmente al interpolar los datos se obtiene una función, la cual no es una función suave de x, entonces para lograr obtener una función suave, se utiliza un término que representa una penalización por falta de suavidad en la función objetivo. La función objetivo sería ahora:

$$\min_{\hat{f}\in J} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2 + \varphi\left(\hat{f}\right) \qquad (2.21)$$

Se plantea el método de mínimos cuadrados penalizados, donde J es una clase de funciones suaves, o sea, deben tener una cantidad de derivadas continuas, por ejemplo p, y $\varphi\left(\hat{f}\right)$ es un funcional, $\varphi : J \to \mathrm{R}$, el cual tomará valores más bajos cuando menos suave sea $\hat{f}$.

Para x_i en el intervalo $[a,b]\subseteq\mathbb{R}$ se puede usar como J al espacio de las funciones que cumplen lo siguiente:

$$J = W_2^2[a,b] = \left\{ f : [a,b] \to \mathrm{R} : \int_a^b (f(x))^2\,dx < \infty, \quad existe\ f''(x)\ y\ \int_a^b (f''(x))^2\,dx < \infty \right\}$$

y como funcional de penalización

$$\varphi(f) = \lambda \int_a^b (f''(x))^2\,dx, \quad \lambda > 0$$

El espacio $W_2^2[a,b]$ se denomina "espacio de Sobolev de segundo orden en $[a,b]$".

De este modo el problema (2.21) se describe como

$$\min_{\hat{f}\in W_2^2[a,b]} \left\{ \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2 + \lambda \int_a^b \left(\hat{f}''(x)\right)^2 dx \right\} \qquad (2.22)$$

Este problema se soluciona usando una función spline cúbica con nodos en los valores de la variable independiente $x_1,\ldots, x_n$.

Se define la función $\eta : [a,b] \to \mathrm{R}$ como un spline de grado p, con nodos $w_1,\ldots,w_k$ si se cumple lo siguiente:

1. Los nodos están en [a, b] y $a = w_0 < w_1 <\ldots< w_k < w_{k+1} = b$.
2. Para cada intervalo $[w_j, w_{j+1}]$, $j = 0\ldots,k$, el polinomio η(x) tiene grado p, o un grado menor.
3. En el intervalo [a, b] la función η(x) tiene (p–1) derivadas continuas.

Las funciones splines de grado 3 o cúbicas son las más utilizadas.

Un spline periódico es aquel que verifica η(a) = η(b).

Un spline es natural de grado p, si p es impar, p = 2m – 1, m ≥ 2, y se cumple que

$$\eta^{(m+j)}(a) = \eta^{(m+j)}(b) = 0, \quad j = 0, 1, ..., m-1.$$

Tenemos un spline cúbico natural si p = 3, m = 2, y se verifican las cuatro restricciones siguientes:

$$\eta''(a) = \eta''(b) = 0, \quad \eta'''(a) = \eta'''(b) = 0.$$

Por lo tanto un spline cúbico natural η(x) es lineal en [a,w_1] y [w_k,b]. Además, $\eta''(w_1) = \eta''(w_k) = 0$.

Proposición

Sea η[p: a = w_o, w_1,…, w_k, w_{k+1} = b] el conjunto de splines de grado p con nodos w_1, …, w_k definidos en el intervalo [a, b], se cumple que el espacio vectorial η[p: a = w_o, w_1,…, w_k, w_{k+1} = b] tiene dimensión p + k + 1.

Proposición

Sea N[p: a = w_o, w_1,…, w_k, w_{k+1} = b] el conjunto de splines naturales de grado p con nodos w_1, …, w_k definidos en el intervalo [a, b], el espacio vectorial N[p: a = w_o, w_1,…, w_k, w_{k+1} = b] tiene dimensión k.

Proposición

Existe sólo un spline natural η(x) de grado p con nodos en x_i, i = 1,…, n, que interpola los puntos (xi, yi), n ≥ 2, y los xi en el intervalo [a, b]:

$$\eta(x_i) = y_i, \; i = 1, \ldots, n$$

2.2.2.6. Suavización por Splines

El concepto de suavización por splines es parecido al de regresión por splines descrito anteriormente. La diferencia está en la forma de obtener la curva final, ya que a diferencia de la regresión spline, en la suavización spline no es necesario elegir el número y la posición de los nodos, ya que hay uno en cada observación. Lo que se tiene que escoger es el valor de λ.

Según Eubank (1999) y Green & Silverman (2000), la función $\int_0^1 f''(x)^2 dx$ sirve para medir la suavidad asociada a una función $f \in W_2^2[0,1]$, y que la bondad del ajuste de los datos al modelo, se mide a través de la suma de cuadrados del error $n^{-1}\sum_{i=1}^{n}(y_i - f(x_i))^2$. Por lo tanto se puede utilizar la suma convexa de estas funciones como una medida de la calidad de un estimador de f, dada por

$$(1-q)\,n^{-1}\sum_{i=1}^{n}(y_i - f(x_i))^2 + q\int_0^1 f''(x)^2 dx$$

Con 0 < q < 1.

Haciendo λ = q / (1 – q), el estimador de f se encuentra eligiendo a f_λ que minimice la suma:

$$n^{-1}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\int_0^1 f''(x)^2 dx \qquad (2.23)$$

Sobre las funciones $f \in W_2^2[0,1]$. A este estimador se le llama un estimador spline de f.

De la expresión (2.23) podemos concluir que si λ es muy grande, entonces la estimación de f será super-suavizada; y cuando λ es muy pequeño, se obtiene un estimador que interpola los datos.

Según Eubank (1999) una solución a este problema de optimización es única y está dada por:

$$f_\lambda = \sum_{i=1}^{n} \beta_{\lambda i}\, f_i \tag{2.24}$$

donde $\beta_\lambda = (\beta_{\lambda 1}, \beta_{\lambda 2}, ..., \beta_{\lambda n})^T$ es la única solución con respecto a $c = (c_1, c_2, ..., c_n)^T$ del sistema de ecuaciones

$$(X^T X + n\lambda\Omega)c = X^T y \tag{2.25}$$

donde

$$X = \{f_j(x_i)\}_{i,j=1,2,...,n}, \qquad y = (y_1, y_2, ..., y_n)^T$$

y $\Omega = \left\{ \int_0^1 f_i''(x)\, f_j''(x)\, dx \right\}_{i,j=1,2,...,n}$

Una base del conjunto de splines naturales está formada por las funciones $\{f_j\}_{j=1,...,n}$. Según Sezer (2009) una base recomendable es la de splines cúbicos naturales.

$$f_1(x) = 1$$
$$f_2(x) = x$$
$$f_{j+2}(x) = d_j(x) - d_{n-1}(x), \quad j = 1, 2, ..., n-2 \tag{2.26}$$
$$donde$$
$$d_j(x) = \frac{\left(x - x_j\right)_+^3 - \left(x - x_n\right)_+^3}{x_j - x_n}$$

y la función $(z)_+^3$ es la función truncada:

$$(z)_+^3 = \begin{cases} z^3, & si\ z \geq 0 \\ 0, & si\ z < 0 \end{cases}$$

Entonces, el vector de valores estimados es $f_\lambda = \left(f_\lambda(x_1), f_\lambda(x_2), ..., f_\lambda(x_n)\right)^T = S_\lambda y$, donde se tiene que

$$S_\lambda = X\left(X^T X + n\lambda\Omega\right)^{-1} X^T \tag{2.27}$$

Al estimador f_λ de f dado en 2.24 se le conoce como estimador spline. El parámetro de suavización λ se elige utilizando la validación cruzada generalizada usando la matriz S_λ (2.27).

2.2.2.7. Elección de λ

Sean y_i las observaciones y $f_\lambda(x_i)$ las estimaciones, se denota RSS(λ) a la suma de cuadrados de las diferencias entre las observaciones y las estimaciones:

$$RSS(\lambda) = \sum (y_i - f_\lambda(x_i))^2 \tag{2.28}$$

Sea **f** el vector de los valores de la función f y $\mathbf{f}_\lambda$ el vector de los valores del estimador f_λ, en los puntos de diseño x_i, y también **y** será el vector de respuestas**.**

Usando esta notación tenemos que:

$$RSS(\lambda) = (\mathbf{y} - \mathbf{f}_\lambda)^T (\mathbf{y} - \mathbf{f}_\lambda)$$

$$= \mathbf{y}^T (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}$$

donde $\mathbf{S}_\lambda = (\mathbf{X}_\lambda^T \mathbf{X}_\lambda)^{-1} \mathbf{X}_\lambda^T \mathbf{X}_\lambda$ y $\mathbf{X}_\lambda = \{f_j(x_i)\}_{i=1,2,\ldots,n;\, j=1,2,\ldots,\lambda}$

Definiremos la *pérdida* de f_λ para estimar f:

$$L_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - f_\lambda(x_i))^2 \tag{2.29}$$

También definimos el *riesgo* de f_λ para estimar f como el valor esperado de la pérdida:

$$R_n(\lambda) = E[L_n(\lambda)] \tag{2.30}$$

Algunos autores prefieren llamar *error cuadrático m*edio, denotado MSE(λ), a lo definido como riesgo.

De la definición del riesgo dada en (2.30) tenemos:

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E\left(f(x_i) - f_\lambda(x_i)\right)^2$$

$$= \frac{1}{n} E\left[(\mathbf{f} - \mathbf{f}_\lambda)^T (\mathbf{f} - \mathbf{f}_\lambda)\right] \tag{2.31}$$

$$= \frac{1}{n}\mathbf{f}^T (\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{f} + \frac{1}{n}\sigma^2 tr\left[\mathbf{S}_\lambda^2\right]$$

Además tenemos que:

$$\begin{aligned} E\left[RSS(\lambda)\right] &= \mathbf{f}^T (\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{f} + \sigma^2 tr\left[(\mathbf{I} - \mathbf{S}_\lambda)^2\right] \\ &= \mathbf{f}^T (\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{f} + \sigma^2 tr\left[\mathbf{S}_\lambda^2\right] + n\sigma^2 - 2\sigma^2 tr\left[\mathbf{S}_\lambda\right] \end{aligned} \tag{2.32}$$

Insertando (2.31) en (2.32), tenemos que:

$$E\left[RSS(\lambda)\right] = nR(\lambda) + n\sigma^2 - 2\sigma^2 tr\left[\mathbf{S}_\lambda\right] \tag{2.33}$$

Podemos observar que $\frac{1}{n}RSS$ es un estimador sesgado de R(λ) con sesgo igual a $\sigma^2 - \frac{2}{n}\sigma^2 tr\left[\mathbf{S}_\lambda\right]$. Luego un estimador insesgado del riesgo R(λ) del estimador f_λ sería:

$$\hat{R}_U(\lambda) = \frac{1}{n}RSS(\lambda) - \sigma^2 + \frac{2}{n}\sigma^2 tr\left[\mathbf{S}_\lambda\right] \tag{2.34}$$

Como no conocemos σ^2 podemos usar cualquiera de los estimadores de la varianza que se verán en la sección posterior, entonces un estimador de R(λ) sería:

$$\hat{R}_U(\lambda) = \frac{1}{n} RSS(\lambda) - \hat{\sigma}^2 + \frac{2}{n}\hat{\sigma}^2 tr[\mathbf{S}_\lambda] \tag{2.35}$$

El estimador de (2.35) se conoce como estimador UBRE por las siglas en inglés de *estimador insesgado del riesgo*. Luego podemos calcular este estimador del riesgo para diferentes valores de λ, y elegir el λ adecuado como aquel para el cual este estimador toma el menor valor.

Otro método para hallar el valor adecuado de λ es el llamado *criterio de validación cruzada*.

Supongamos que queremos obtener n nuevas observaciones de Y que se supone pueden ser modelados de la misma forma que los datos originales. Sea Y_N el vector de las nuevas observaciones. Ahora digamos que queremos predecir "y" usando f_λ, definimos el *riesgo de predicción* P(λ), como sigue:

$$P(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E[y - f]^2 \tag{2.36}$$

La definición de nuevas observaciones permite afirmar que:

$$P(\lambda) = \sigma^2 + R(\lambda) \tag{2.37}$$

Una forma de obtener nuevas observaciones consiste en dividir el conjunto de n observaciones en n submuestras de tamaño n – 1, dejando fuera de la submuestra una observación diferente cada vez. Sea $f_{\lambda(i)}$ la estimación de $f_i = f(x_i)$ obtenida al suprimir de la muestra la observación i,

entonces la observación y_i sería una observación adicional, luego un estimador de P(λ) sería:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_{\lambda(i)}\right)^2 \tag{2.38}$$

La expresión anterior (2.38) se conoce como el criterio de validación cruzada.

El cálculo de CV en 2.38 utiliza una gran carga de procesamiento computacional, Eukbank (1999), propone la siguiente expresión llamada criterio de *validación cruzada generalizada*:

$$GCV(\lambda) = \frac{n^{-1}RSS(\lambda)}{\left(n^{-1}tr[\mathbf{I} - \mathbf{S}_{\boldsymbol{\lambda}}]\right)^2} \tag{2.39}$$

La palabra generalizada no se refiere a que esta expresión generaliza la primera, en realidad se trata de criterios diferentes para estimar el riesgo de predicción P(λ).

Wahba (1990) indica que el uso de GCV es un buen método de selección de λ, demostrando que si $n^{-1}tr[\mathbf{S}_{\boldsymbol{\lambda}}] < 1$, la diferencia entre E[GCV(λ)] y P(λ) relativa al tamaño de R(λ) será pequeña, más aún si el tamaño de muestra es grande.

2.2.2.8. Estimación de la varianza

Primeramente se podría utilizar el estimador de la varianza del análisis de regresión lineal dado por

$$\sigma_{\lambda}^{2} = \frac{SCE_{\lambda}}{(n-\lambda)} \tag{2.40}$$

donde la suma de cuadrados de los residuales SCE_{λ} se define como

$$SCE_{\lambda} = \sum_{i=1}^{n} \left(y_i - f_{\lambda}\left(x_i\right)\right)^2 \tag{2.41}$$

El estimador de la varianza dado en (2.40) en general no es insesgado, y su sesgo está dado por:

$$B\left(\sigma_{\lambda}^{2}\right) = \frac{\mathbf{f}^{T}\left(\mathbf{I} - \mathbf{S}_{\lambda}\right)\mathbf{f}}{n-\lambda} \tag{2.42}$$

con $\mathbf{f} = (f(x_1),\ldots,f(x_n))^{\mathbf{T}}$ y $\mathbf{S}_{\lambda} = \mathbf{X}_{\lambda}\left(\mathbf{X}_{\lambda}^{T}\mathbf{X}_{\lambda}\right)^{-1}\mathbf{X}_{\lambda}^{T}$.

Este sesgo tiene el efecto de sobreestimar el parámetro de varianza.

Como se puede observar en regresión no paramétrica existen una gran cantidad de estimadores de σ^2, se puede considerar que existen dos grupos. En el primero de ellos, estimadores que dependen del parámetro de suavización, se realiza la estimación de la varianza a partir de la suma de cuadrados de errores obtenidos de ajustar un modelo no paramétrico de f, por ejemplo suavización Kernel o regresión por Splines, y en el segundo grupo se encuentran los estimadores que utilizan diferencias, los cuales

usan las respuestas y_i correspondientes a una vecindad de x, estos estimadores no dependen explícitamente del parámetro de suavización. El número de observaciones utilizadas en el cálculo de los residuales determina el orden de los estimadores de diferencias.

Por otra parte una estimación de la varianza que no dependa de λ, también la haría independiente del estimador de f. John Rice en 1984 propuso un estimador al que llamaremos *estimador de Rice,* definido de la siguiente manera

$$\sigma_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} \left(y_i - y_{i-1} \right)^2 \tag{2.43}$$

Posteriormente en 1986, Gasser, Sroka & Jennen-Steinmetz definieron unos pseudo-residuales dados a continuación:

$$\tilde{\varepsilon}_i = \frac{\left(y_i - C_i y_{i-1} - D_i y_{i+1} \right)}{\left(1 + C_i^2 + D_i^2 \right)^{1/2}}, \quad i = 2,\ldots,n-1 \tag{2.44}$$

donde

$$C_i = \frac{\left(x_{i+1} - x_i \right)}{\left(x_{i+1} - x_{i-1} \right)}$$

$$D_i = \frac{\left(x_i - x_{i-1} \right)}{\left(x_{i+1} - x_{i-1} \right)}$$

Basándonos en estos pseudo-residuales el estimador de la varianza denotado por σ_{GSJS}^2, está dado a continuación

$$\sigma^2_{GSJS} = \frac{1}{n-2}\sum_{i=2}^{n-1}\tilde{\varepsilon}_i^{\,2} \tag{2.45}$$

Luego en 1990, Hall, Kay y Titterington proponen otro estimador basado en diferencias sucesivas en forma general. Lo llamaremos estimador HKT y lo denotaremos por σ^2_{HKT}.

Sean m_1 y m_2 dos enteros no negativos tales que $m_1 + m_2 = r$. El estimador σ^2_{HKT} se basa en sucesiones de diferencias de orden r $\{d_k\}_{k=-m_1}^{m_2}$, que cumplen las siguientes condiciones

$$\sum_{k=-m_1}^{m_2} d_k = 0 \quad y \quad \sum_{k=-m_1}^{m_2} d_k^2 = 1. \qquad \text{donde } d_k \neq 0$$

Se asume que $d_k = 0$ para $k < -m_1$, $k > m_2$ y $d_{-m_1} d_{m_2} \neq 0.$

Luego el estimador σ^2_{HKT} se define como:

$$\sigma^2_{HKT} = \frac{1}{n-r}\sum_{l=m_1+1}^{n-m_2}\left(\sum_{k=-m_1}^{m_2} d_k\, y_{k+l}\right)^2 \tag{2.46}$$

Para $m_1 = 0$ y $m_2 = r$ tenemos:

$$\sigma^2_{HKTr} = \frac{1}{n-r}\sum_{l=1}^{n-r}\left(\sum_{k=0}^{r} d_k\, y_{k+l}\right)^2 \tag{2.47}$$

Todos los estimadores de la varianza vistos están pensados bajo el modelo homocedástico.

2.2.2.9 Modelos Aditivos Generalizados

En el caso de la regresión múltiple, cuando tenemos una variable dependiente Y, y p variables independientes X_i, i = 1,…,p, al tratar de aplicar un método de suavización, se presenta el problema de la dimensionalidad, llamado también como la maldición de la dimensionalidad. Cuando el número de variables independientes se hace mayor, entonces vecindades locales, contienen cada vez menos puntos cercanos al punto para el cual se quiere estimar la función de regresión, necesitando entonces aumentar el tamaño de la muestra. Para tratar de solucionar esta situación se tiene los Modelo Aditivos Generalizados (GAM).

En este caso el modelo de regresión toma la forma

$$Y = \sum_{i=1}^{p} f_i(X_i) + \varepsilon \tag{2.48}$$

Las funciones f_i, univariadas, se obtienen aplicando el procedimiento iterativo llamado backfitting.

Algoritmo:

1) Se empieza definiendo las funciones $f_j^{(0)} = 1$, j = 1,…,p.

2) En la i-ésima iteración se obtienen las estimaciones $f_j^{(i+1)}$, aplicando cualquier método de suavización a la función $Y - \sum_{k \neq j} f_k^{(i)}(X_k)$, j =1,…,p.

3) Dada una constante δ de tolerancia, verificar si $\left| f_j^{(i+1)} - f_j^{i} \right| < \delta$. Si se cumple la condición se detiene el proceso iterativo, de no ser así, se repite el paso 2.

2.2.2.10 Regresión de búsqueda de proyección (Regresión por Projection Pursuit), PPR

En este caso tenemos puntos en un espacio p-dimensional, entonces se busca reducir la dimension de los datos proyectándolos en un espacio de menor dimension. Para esto podemos utilizar el Análisis de Componentes Principales (ACP), mediante el cual las p variables se pueden reducir a 1 o 2 componentes principales, conservando la máxima varianza en los datos proyectados.

Se tiene el vector **X** con p componentes y una variable Y. Sea w_m, m = 1,2…,M, el vector de parámetros desconocidos. El modelo de regression por Projection Pursuit está dado por

$$f(x) = \sum_{m=1}^{M} f_m\left(w_m^T \mathrm{X}\right)$$

Las funciones f_m son estimadas usando cualquier método de suavización, a través de las direcciones w_m.

Se ajusta un modelo PPR a los datos ($\mathbf{x}_i$,y_i), i = 1,…,M, minimizando la siguiente función:

$$\sum\left[y_i - \sum f_m\left(w_m^T\right)\right]^2$$

2.2.2.11 Regresión por Árboles Cart

La superficie de regresión se estima usando el modelo aditivo

$$f(x) = \sum_{i=1}^{n} c_i I_{Ni}(x)$$

donde

- c_i son constantes
- I es la función indicadora, tal que $I_{Ni}(x) = 1,$ si $x \in N_i$.
- Los Ni son hiperrectángulos disjuntos con lados paralelos a los ejes coordenados. Los hiperrectángulos se obtienen por partición recursivas y pueden ser representados por árboles.

CAPÍTULO 3: Metodología

3.1 Tipo y Diseño de Investigación

Se llevó a cabo una investigación de tipo básica.

Se realizó un estudio de simulación, con el cual se obtuvieron valores de un modelo de regresión, el cual está planteado a continuación:

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1,...,n$$

Los valores de y_i son las respuestas, f es la función de regresión poblacional, x_i la covariable y ε_i los errores aleatorios.
Se cumplen las siguientes condiciones:

- El diseño es fijo, y los valores de x_i se encuentran ordenados equidistantemente en el intervalo [0,1], considerando un solo valor de y_i en cada punto de diseño, lo que significa que no se tienen medidas repetidas.
- Los valores de los errores aleatorios son independientes y tienen la misma distribución, con $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma^2$.
- La función f es continua y tiene al menos las dos primeras derivadas.

El procedimiento de simulación se realizó para diferentes casos. Se consideraron tres tipos de funciones de regresión f, una sin oscilaciones, otra con un número de oscilaciones bajo y la última con un número de oscilaciones alto. También se utilizarán diferentes distribuciones para los errores.

Las funciones de regresión poblacionales seleccionadas fueron:

- $4 \sin(0.5 \pi x_i) - 2$, no tiene oscilaciones.
- $2 \sin(3 \pi x_i)$, tiene un número bajo de oscilaciones.
- $2 \sin(7 \pi x_i)$, tiene un número alto de oscilaciones.

Varianza de los errores: $\sigma^2 = 0.5$ (baja variabilidad), $\sigma^2 = 1$ (alta variabilidad).
Distribución de ε_i:

- $N(0,\sigma^2)$, distribución simétrica.
- Para la distribución asimétrica hacia la derecha se consideró la distribución semi-normal: $\left| N(0,1) \right| - \left(\frac{2}{\pi} \right)^{1/2}$.

- Para la distribución asimétrica hacia la izquierda se consideró la distribución semi-normal: $\left(\frac{2}{\pi}\right)^{1/2} - \left| N(0,1) \right|$.

Los estimadores no paramétricos que se compararon son el LOESS y la suavización por Splines.

Para comparar los estimadores en cada caso se utilizó el error cuadrático medio.

Se realizó este trabajo utilizando el programa estadístico R.

3.2 Tamaño de muestra

Para las diferentes funciones de regresión indicadas y las diferentes distribuciones del error del modelo, se generaron muestras de tamaño 100, 200 y 300.

Para poder obtener resultados confiables, se realizaron 1000 repeticiones para cada simulación.

En la mayoría de los trabajos revisados se utilizó como mínimo un tamaño de muestra de 50. Por la complejidad de los modelos de regresión no paramétrica no se encontró una fórmula de donde se pueda deducir el tamaño de muestra mínimo. En todo caso para poder encontrar el tamaño de muestra mínimo necesario para aplicar estos métodos, se recomienda realizar un trabajo de simulación posteriormente.

CAPÍTULO 4: Resultados y discusión

4.1 Análisis e interpretación de la información

Se aplican técnicas estadísticas que permiten comparar las diferentes medidas calculadas para escoger el mejor modelo en cada caso.

A continuación se presentan las funciones de regresión poblacionales utilizadas para realizar la simulación.

En la figura 4 el caso (a) presenta una función de regresión sin oscilaciones, el caso (b), presenta la función de regresión con pocas

oscilaciones y el caso (c) presenta la función de regresión con muchas oscilaciones.

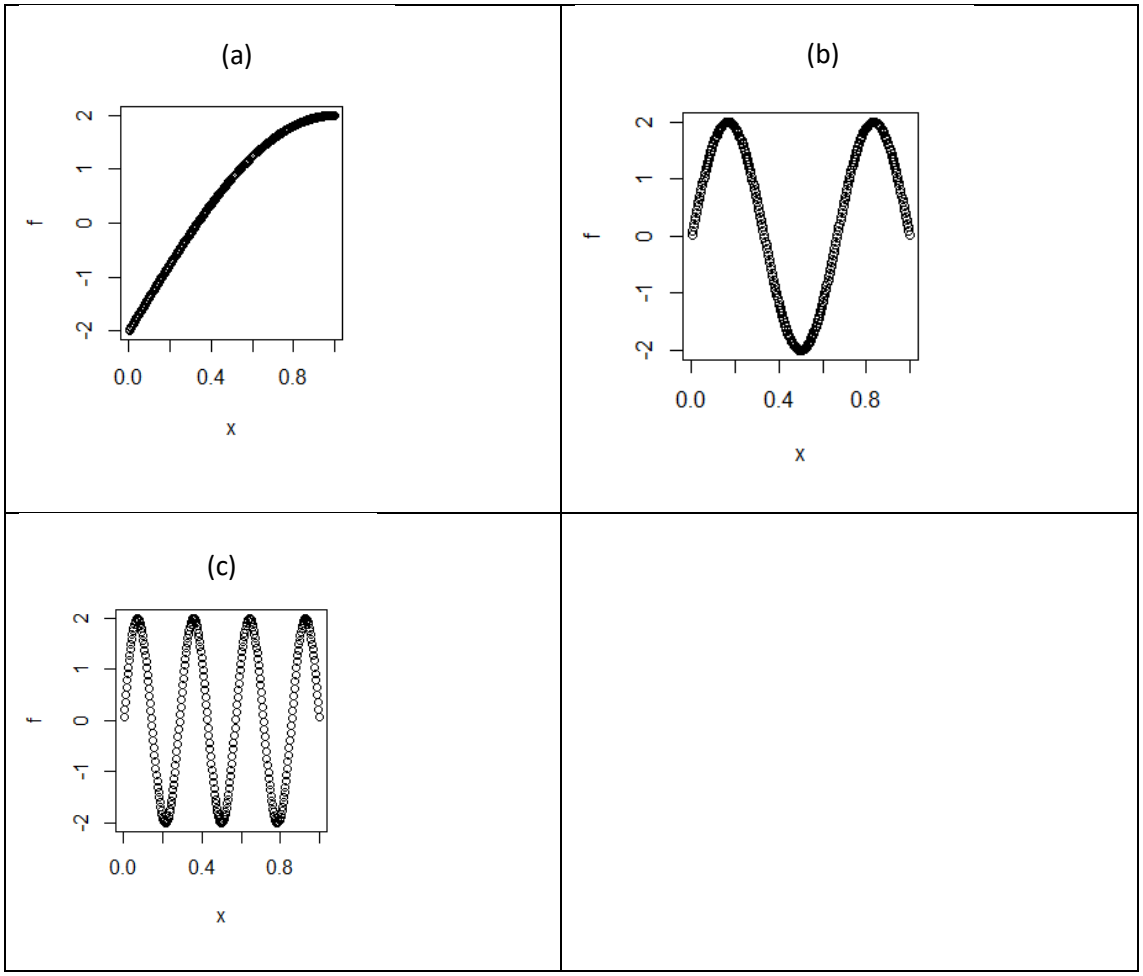

Figura 4. En el caso (a) se presenta el gráfico de los datos generados sin oscilaciones, en el caso (b) los datos son generados con pocas oscilaciones y en caso (c) los datos son generados con muchas oscilaciones.

A continuación en la figura 5 se muestran los gráficos de caja de los errores generados para el proceso de simulación desarrollado. El caso (a) presenta a los errores generados con distribución simétrica, en el caso (b), los errores se generaron con distribución asimétrica a la izquierda y finalmente, en el caso (c), los errores se generaron con distribución asimétrica a la derecha.

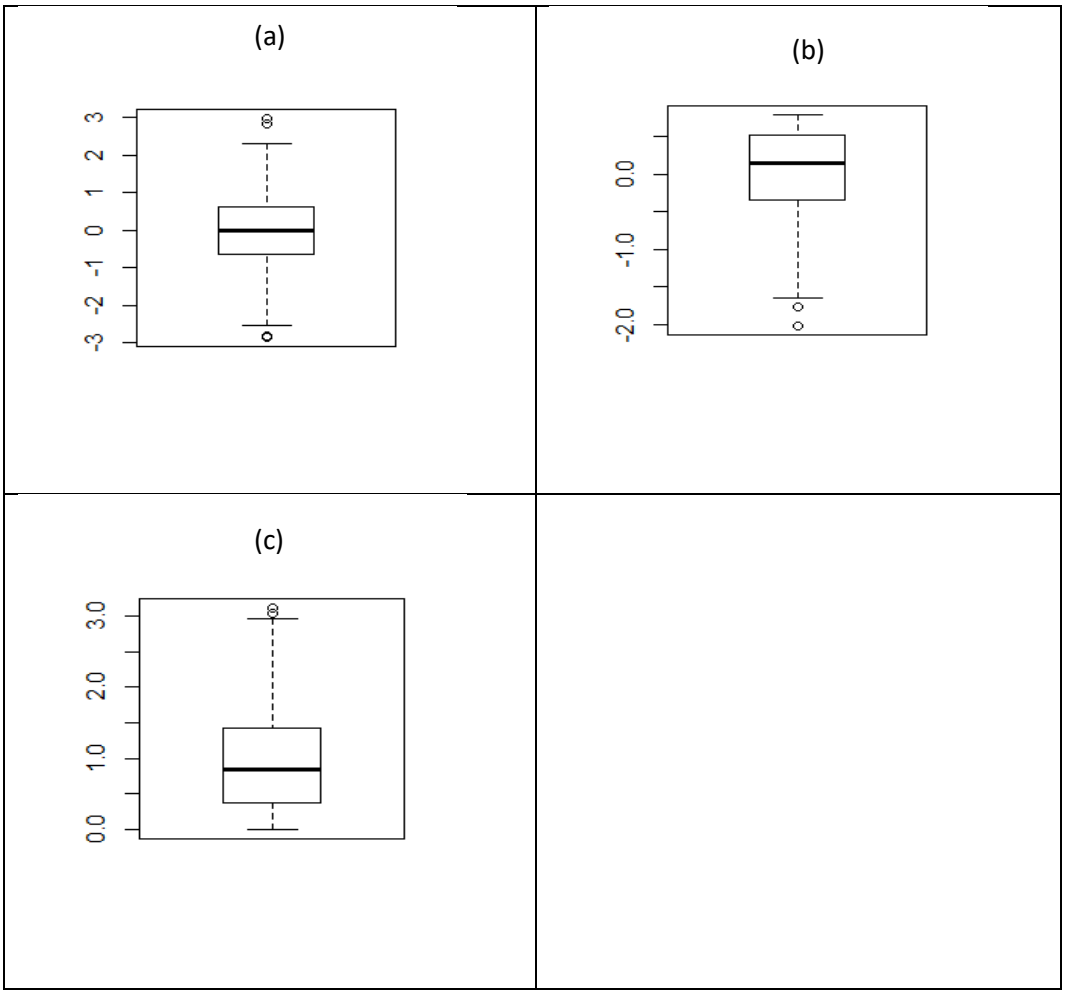

Figura 5. Distribución de los errores generados de los modelos de regresión. En el caso (a) se presentan errores con distribución simétrica, en el caso (b) la distribución es asimétrica a la izquierda y en caso (c) la distribución es asimétrica a la derecha.

Se llevó a cabo cada una de las simulaciones, considerando una función de regresión poblacional, una distribución de los errores, simétrica, asimétrica a la derecha y asimétrica a la izquierda, varianzas de los errores iguales a 0.5 y 1, para un tamaño de muestra específico, 100, 200 y 300, y un método de regresión no paramétrica, LOESS y suavización Spline.

Cada simulación se repitió 1000 veces, y se obtuvo el ECM para las repeticiones consideradas, en las tablas siguientes se presenta el promedio de los ECM correspondientes a las 1000 simulaciones.

Tabla 1

Error cuadrático medio de los estimadores cuando se utiliza una función de regresión simétrica sin oscilaciones

	n = 100 $\sigma^2 = 0.5$	**n = 100 $\sigma^2 = 1$**	**n = 200 $\sigma^2 = 0.5$**	**n = 200 $\sigma^2 = 1$**	**n = 300 $\sigma^2 = 0.5$**	**n = 300 $\sigma^2 = 1$**
Suavización Spline	0.02546	0.05213	0.01310	0.02676	0.00876	0.01754
LOESS	0.04842	0.09616	0.02429	0.04824	0.01635	0.03236

En la tabla 1 podemos observar que cuando la función de regresión es simétrica sin oscilaciones los ECM del método LOESS, en todos los casos, son aproximadamente el doble que ECM de la suavización Spline. Considerando las varianzas, sin importar el método, tenemos que para los casos de poca variabilidad, $\sigma^2 = 0.5$, los ECM son más o menos la mitad de los ECM de los casos cuando la variabilidad es mayor.

Tabla 2

Error cuadrático medio de los estimadores cuando se utiliza una función de regresión simétrica con pocas oscilaciones

	n = 100 $\sigma^2 = 0.5$	**n = 100 $\sigma^2 = 1$**	**n = 200 $\sigma^2 = 0.5$**	**n = 200 $\sigma^2 = 1$**	**n = 300 $\sigma^2 = 0.5$**	**n = 300 $\sigma^2 = 1$**
Suavización Spline	0.04307	0.07935	0.02269	0.04214	0.01565	0.02924
LOESS	0.04961	0.09653	0.02497	0.04707	0.01677	0.03295

En la tabla 2 cuando la función de regresión es simétrica con pocas oscilaciones, podemos observar que los ECM son similares para ambos métodos, sin importar el tamaño de la muestra. Considerando las varianzas, sin importar el método, tenemos que para los casos de poca variabilidad, $\sigma^2 = 0.5$, los ECM son más o menos la mitad de los ECM de los casos cuando la variabilidad es mayor.

Tabla 3

Error cuadrático medio de los estimadores cuando se utiliza una función de regresión simétrica con muchas oscilaciones

	n = 100 $\sigma^2 = 0.5$	**n = 100 $\sigma^2 = 1$**	**n = 200 $\sigma^2 = 0.5$**	**n = 200 $\sigma^2 = 1$**	**n = 300 $\sigma^2 = 0.5$**	**n = 300 $\sigma^2 = 1$**
Suavización Spline	0.07862	0.14532	0.04077	0.07722	0.02156	0.02156
LOESS	0.13559	0.18337	0.11120	0.13751	0.10170	0.11808

En la tabla 3, cuando la función de regresión es simétrica con muchas oscilaciones, La suavización Spline presenta los menores ECM en todos los casos, pero cuando el tamaño de muestra aumenta la diferencia aumenta. Estos ECM son muchos mayores que los obtenidos cuando la función de regresión es simétrica.

Tabla 4

Error cuadrático medio de los estimadores cuando se utiliza una función de regresión asimétrica a la derecha sin oscilaciones

	n = 100 $\sigma^2 = 0.5$	**n = 100 $\sigma^2 = 1$**	**n = 200 $\sigma^2 = 0.5$**	**n = 200 $\sigma^2 = 1$**	**n = 300 $\sigma^2 = 0.5$**	**n = 300 $\sigma^2 = 1$**
Suavización Spline	0.01895	0.01869	0.00971	0.01006	0.00648	0.00698

LOESS	0.03573	0.03509	0.01808	0.01783	0.01173	0.01194

En la tabla 4, podemos observar que cuando la distribución es asimétrica a la derecha sin oscilaciones, los ECM de la distribución Spline son menores que los ECM del LOESS, pero en todos los casos estos ECM son menores que cuando la función de regresión era simétrica sin oscilaciones.

Tabla 5

Error cuadrático medio de los estimadores cuando se utiliza una función de regresión asimétrica a la derecha con pocas oscilaciones

	n = 100 $\sigma^2 = 0.5$	**n = 100 $\sigma^2 = 1$**	**n = 200 $\sigma^2 = 0.5$**	**n = 200 $\sigma^2 = 1$**	**n = 300 $\sigma^2 = 0.5$**	**n = 300 $\sigma^2 = 1$**
Suavización Spline	0.03288	0.03258	0.01743	0.01744	0.01219	0.01192
LOESS	0.03611	0.03628	0.01769	0.01821	0.01215	0.01227

En la tabla 5, cuando la distribución de regresión es asimétrica a la derecha con pocas oscilaciones, podemos observar que los ECM son muy parecidos con ambos modelos de regresión. También se observa que los ECM son similares para los casos de poca variabilidad y mucha variabilidad de los errores.

Tabla 6

Error cuadrático medio de los estimadores cuando se utiliza una función de regresión asimétrica a la derecha con muchas oscilaciones

	n = 100 $\sigma^2 = 0.5$	**n = 100 $\sigma^2 = 1$**	**n = 200 $\sigma^2 = 0.5$**	**n = 200 $\sigma^2 = 1$**	**n = 300 $\sigma^2 = 0.5$**	**n = 300 $\sigma^2 = 1$**
Suavización Spline	0.05875	0.05750	0.03054	0.03138	0.02130	0.02195

LOESS	0.12293	0.12202	0.10599	0.10477	0.09624	0.09714

En la tabla 6, cuando la función de regresión es asimétrica con muchas oscilaciones, podemos observar que la suavización Spline presenta los menores ECM en todos los casos, pero cuando el tamaño de muestra aumenta la diferencia aumenta. Sin considerar el tamaño de muestra podemos observar que los ECM son similares tanto para poca variabilidad como para mucha variabilidad de los errores.

Tabla 7

Error cuadrático medio de los estimadores cuando se utiliza una función de regresión asimétrica a la izquierda sin oscilaciones

	n = 100 $\sigma^2 = 0.5$	**n = 100 $\sigma^2 = 1$**	**n = 200 $\sigma^2 = 0.5$**	**n = 200 $\sigma^2 = 1$**	**n = 300 $\sigma^2 = 0.5$**	**n = 300 $\sigma^2 = 1$**
Suavización Spline	0.01873	0.01902	0.00951	0.00991	0.00698	0.00692
LOESS	0.03447	0.03561	0.01782	0.01777	0.01195	0.01187

En la tabla 7 se puede observar que los ECM de la suavización Spline son menores, y que todos los valores, sin importar el método, ni la varianza ni el tamaño de la muestra, son menores a los ECM cuando la distribución es simétrica sin oscilaciones.

Tabla 8

Error cuadrático medio de los estimadores cuando se utiliza una función de regresión asimétrica a la izquierda con pocas oscilaciones

	n = 100 $\sigma^2 = 0.5$	**n = 100 $\sigma^2 = 1$**	**n = 200 $\sigma^2 = 0.5$**	**n = 200 $\sigma^2 = 1$**	**n = 300 $\sigma^2 = 0.5$**	**n = 300 $\sigma^2 = 1$**
Suavización	0.03355	0.03416	0.01683	0.01680	0.01195	0.01188

Spline						
LOESS	0.03628	0.03496	0.01800	0.01853	0.01208	0.01206

En la tabla 8, cuando la distribución es asimétrica hacia la izquierda con pocas oscilaciones, podemos observar que los ECM son similares para ambos métodos de estimación. No hay diferencia entre los ECM cuando la variabilidad de los errores aumenta de 0.5 a 1 y se mantiene fijo el tamaño de la muestra.

Tabla 9

Error cuadrático medio de los estimadores cuando se utiliza una función de regresión asimétrica a la izquierda con muchas oscilaciones

	n = 100 $\sigma^2 = 0.5$	**n = 100 $\sigma^2 = 1$**	**n = 200 $\sigma^2 = 0.5$**	**n = 200 $\sigma^2 = 1$**	**n = 300 $\sigma^2 = 0.5$**	**n = 300 $\sigma^2 = 1$**
Suavización Spline	0.05769	0.05740	0.03108	0.03081	0.02152	0.02120
LOESS	0.12167	0.12230	0.10483	0.10403	0.09599	0.09663

La tabla 9, cuando se utiliza una función de regresión asimétrica a la izquierda con muchas oscilaciones, podemos observar que los ECM de la suavización Spline son menores que el LOESS, pero a medida que el tamaño de muestra aumenta la diferencia se hace mayor. Cuando se comparan los ECM para los dos casos de variabilidad de los errores observamos que los ECM son similares.

CAPÍTULO 5: Conclusiones y recomendaciones

5.1 Conclusiones

- Se puede concluir que los métodos de regresión Loess y Spline si son adecuados para la predicción de los valores de la variable respuesta.
- Se puede concluir que en todos los casos, como es de esperar, los errores cuadráticos medios disminuyen al aumentar el tamaño de la muestra.
- También se puede concluir que cuando la función de regresión no tiene oscilaciones o cuando tiene muchas oscilaciones los ECM son menores en el caso de la suavización Spline, siendo este método de regresión no paramétrica el adecuado en estos casos.
- También se ha podido observar que cuando la función de regresión tiene pocas oscilaciones, tanto en el caso de regresión simétrica como asimétrica, sin importar que la función de regresión mucha o poca varianza, los ECM son similares en ambos métodos de regresión no paramétrica. Por lo tanto la suavización Spline y el LOESS tienen un comportamiento similar.
- Podemos observar que cuando la distribución es simétrica y los errores tienen poca variabilidad los ECM son aproximadamente la mitad de lo que son cuando la variabilidad aumenta, sin importar si la función de regresión tenga o no oscilaciones.
- Cuando la función de regresión es asimétrica tanto a la derecha como a la izquierda, el valor del ECM permanece inalterable, cuando se aumenta la variabilidad de los errores.

5.2 Recomendaciones

Se recomienda utilizar la suavización Spline en lugar de LOESS cuando la función de regresión no tiene oscilaciones o tiene muchas oscilaciones, tanto para funciones de regresión simétricas o asimétricas. En el caso de que la función de regresión tenga pocas oscilaciones, ambas estimaciones, suavización Spline y LOESS son recomendables ya que tienen un comportamiento similar.

Se recomienda extender el estudio realizado tomando en cuenta los casos en que se presentan medidas repetidas, y los casos en los cuales los valores de la variable predictora no son equidistantes.

A Anexos

A. 1

Programa R utilizado para obtener el Regresorgrama de la figura 1:

```
x<-vector('numeric')
y<-vector('numeric')
z<- vector('numeric')
x<-1:100
for(i in 1:100)

{y[i]<-5*x[i]+rnorm(1,50,100)}
c=5
n<-length(x)
x<-sort(x)
y<-y[order(x)]
z<-x[1]
intervalos<-floor(n/c)
ymedias<-rep(0,c)
for(j in 1:c)
  {indicador<-((j-1)*intervalos+1):(j*intervalos)
  if(j<c)
    z<-c(z,x[j*intervalos])
  if(j==c)
  {indicador<-((j-1)*intervalos+1):n
  z<-c(z,x[n])
  }
  ymedias[j]<-mean(y[indicador])
}
z<-c(z,z[2:c])
z<-sort(z)
```

```
  ymedias1<-rep(ymedias,each=2)
  plot(x,y)
  #lines(z,ymedias1)
  title("Regresorgrama")
  cat("\las medias de y en cada subintervalo son:\n")
  ymedias

lines(z,ymedias1)
```

A. 2

Programa R utilizado para obtener el gráfico de Medias Móviles de la figura 2:

```
x<-vector('numeric')
y<-vector('numeric')
x<-1:100
for(i in 1:100)
{y[i]<-5*x[i]+rnorm(1,50,100)}
n<-length(x)
r<-rep(0,n)
for(i in 1:n)
{ind1<-max(i-k,1)
ind2<-min(i+k,n)
z<-y[ind1:ind2]
r[i]<-mean(z)
}
plot(x,y)
lines(sort(x),r,type="l")
title("Running means")
```

A. 3

Programa R para Suavización Splines modelo simétrico sin oscilaciones:

```
n<-300
x<-numeric(0)
```

```
f<-numeric(0)
y<-numeric(0)
ECM<- numeric(0)
for (j in 1:1000)
{for (i in 1:n)
x[i]<-(2*i-1)/(2*n)
x
sigma<-0.7071
for (i in 1:n)
{f[i]<- 4*sin(0.5*pi*x[i])-2
y[i]<-f[i]+rnorm(1,0,sigma)}
y
modelo_sp<-smooth.spline(y~x,cv = TRUE)
modelo_sp$lambda
fhat<-predict(modelo_sp)
dif<-fhat$y-f
dif
ECM[j]<-sum(dif**2)/n}
ECM
ECMPROM<-mean(ECM)
ECMPROM
```

A. 4

Programa R para Suavización Splines modelo simétrico con pocas oscilaciones:

```
n<-300
x<-numeric(0)
f<-numeric(0)
```

```
y<-numeric(0)
ECM<- numeric(0)
for (j in 1:1000)
{for (i in 1:n)
x[i]<-(2*i-1)/(2*n)
x
sigma<-0.7071
for (i in 1:n)
{f[i]<- 2*sin(3*pi*x[i])
y[i]<-f[i]+rnorm(1,0,sigma)}
y
modelo_sp<-smooth.spline(y~x,cv = TRUE)
modelo_sp$lambda
fhat<-predict(modelo_sp)
dif<-fhat$y-f
dif
ECM[j]<-sum(dif**2)/n}
ECM
ECMPROM<-mean(ECM)
ECMPROM
```

A. 5

Programa R para Suavización Splines modelo simétrico con muchas oscilaciones:

```
n<-300
x<-numeric(0)
f<-numeric(0)
```

```
y<-numeric(0)
ECM<- numeric(0)
for (j in 1:1000)
{for (i in 1:n)
x[i]<-(2*i-1)/(2*n)
x
sigma<-0.7071
for (i in 1:n)
{f[i]<- 2*sin(7*pi*x[i])
y[i]<-f[i]+rnorm(1,0,sigma)}
y
modelo_sp<-smooth.spline(y~x,cv = TRUE)
modelo_sp$lambda
fhat<-predict(modelo_sp)
dif<-fhat$y-f
dif
ECM[j]<-sum(dif**2)/n}
ECM
ECMPROM<-mean(ECM)
ECMPROM
```

A. 6

Programa R para Suavización Splines modelo asimétrico a la izquierda con muchas oscilaciones:

```
n<-300
x<-numeric(0)
f<-numeric(0)
```

```
y<-numeric(0)
ECM<- numeric(0)
for (j in 1:1000)
{for (i in 1:n)
x[i]<-(2*i-1)/(2*n)
x
sigma<-0.7071
for (i in 1:n)
{f[i]<- 2*sin(7*pi*x[i])
y[i]<-f[i]+((2/pi)^0.5)-abs(rnorm(1,0,1))}
y
modelo_sp<-smooth.spline(y~x,cv = TRUE)
modelo_sp$lambda
fhat<-predict(modelo_sp)
dif<-fhat$y-f
dif
ECM[j]<-sum(dif**2)/n}
ECM
ECMPROM<-mean(ECM)
ECMPROM
```

A. 7

Programa R para Suavización Splines modelo asimétrico a la derecha con muchas oscilaciones:

```
n<-300
x<-numeric(0)
f<-numeric(0)
```

```
y<-numeric(0)
ECM<- numeric(0)
for (j in 1:1000)
{for (i in 1:n)
x[i]<-(2*i-1)/(2*n)
x
sigma<-0.7071
for (i in 1:n)
{f[i]<- 2*sin(7*pi*x[i])
y[i]<-f[i]+ abs(rnorm(1,0,1))}-((2/pi)^0.5)
y
modelo_sp<-smooth.spline(y~x,cv = TRUE)
modelo_sp$lambda
fhat<-predict(modelo_sp)
dif<-fhat$y-f
dif
ECM[j]<-sum(dif**2)/n}
ECM
ECMPROM<-mean(ECM)
ECMPROM
```

A. 8

Programa en R para aplicar Loess al modelo simétrico sin oscilaciones:

```
n<-100
x<-numeric(0)
f<-numeric(0)
y<-numeric(0)
```

```
ECM<- numeric(0)
for (j in 1:1000)
{for (i in 1:n)
  x[i]<-(2*i-1)/(2*n)
x
sigma<-1
for (i in 1:n)
{f[i]<- 4*sin(0.5*pi*x[i])-2
y[i]<-f[i]+rnorm(1,0,sigma)}
y
modelo_loess<-loess(y~x,span = 0.3)
fhat<-predict(modelo_loess)
dif<-fhat-f
dif
ECM[j]<-sum(dif**2)/n}
ECM
ECMPROM<-mean(ECM)
```

Referencias Bibliográficas

1. Acuña E. (2007), *Regresión no paramétrica.* Capítulo 9, Recuperado de https://www.academia.edu/37271639/ANALISIS_DE_REGRESION.
2. Benedettti, G. (1975), *Kernel estimation of regresión functions, Proceedings in Computer Science and Statistics: 8th anual symposium on the interface* pp. 405-412.

3. Cleveland, W. (1979), *Robust locally weighted regression and smoothing scatterplots,* Journal of the American Statistical Association **74**, 829-836.
4. Delicado, P. (2008), *Curso de Modelos no Paramétricos,* UPC.
5. Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing,* 2ª ed., Marcel Dekker, New York, NY.
6. Ferraty, F. & Nuñez Anton, V. & Vieu. P. (2001), *Regresión no Paramétrica: Desde la Dimensión Uno Hasta la Dimensión Infinita*, UPV/EHU.
7. Gasser, T., Müller, H. (1979), *Smoothing Techniques for Curve Estimation,* Springer, Heidelberg.
8. Graybill, F. A. (1976), *Theory and Application of the Linear Model,* Wadsworth & Brooks, Pacific Grove, California.
9. Green P., Silverman, B. (2000), *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach,* Chapman &Hall/CRC, Boca Raton, FL.
10. Meza L. (2013), *Regresión no paramétrica utilizando Spline para la suavización de la estructura de la mortalidad en el Perú,* (Tesis de pregrado). Universidad Nacional Mayor de San Marcos. Perú.
11. Nadaraya, E. A. A. & Seckler, B. T. (1964), '*On estimación regression'*, *Theory of Probability and its Aplicatons (Transl of Teorija Verojatnosteri i ee Primenenija)* **9,** pp 141 – 142**.**
12. Olaya, J. (2012). *Métodos de Regression No Paramétrica,* Ed. Universidad del Valle, Cali.
13. Pereira, L. A., Paz, M. C. & Olaya, J. (2007), '*Estimación de la varianza en regresión no-paramétrica: El efecto de poseer múltiples observaciones por punto de diseño, in'17mo. Simposio de Estadística'*, Universidad Nacional de Colombia.
14. Priestley, M., Chao, M. (1972), *Non-parametric function fitting,* Journal of the Royal Statistical Society, *Series B* **34***, 385-392.*

15. R Development Core Team (2011), *R: A Language and Environment for Statistical Computing,* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
    Recuperado de http://www.R-project.org/
16. Sezer, A. (2009), *Asseing the quality of the natural cubic spline approximation,* Proceedings of the 8th WSEAS International Conference on SYSTEM SCIENCE and SIMULATION in ENGINEERING pp. 186-190.
17. Searle, S. R. (1971), *Linear Model*, John Wiley & Sons, New York, NY.
18. Wang. J.-L. (2003), '*Nonparametric regression analysis of longitudinal data'*.
    Recuperado de http://www.stst.ucdavis.edu/ ˜wang/paper/EOB3.pdf
19. Wahba, G. (1990), *Spline Models for Observational data,* CBMS-NSF Series, SIAM.
20. Watson, G. (1964), '*Smooth regression analysis'*, Sankhya, series A 26, pp 359-372.